Title
Marchenko-Pastur law with relaxed independence conditions

Permalink
https://escholarship.org/uc/item/91v5d4j0

Authors
Bryson, Jennifer
Vershynin, Roman
Zhao, Hongkai

Publication Date
2019-12-29

MARCHENKO-PASTUR LAW WITH RELAXED INDEPENDENCE CONDITIONS

JENNIFER BRYSON, ROMAN VERSHYNIN AND HONGKAI ZHAO

Abstract. We prove the Marchenko-Pastur law for the eigenvalues of sample covariance matrices in two new situations where the data does not have independent coordinates. In the first scenario – the block-independent model – the coordinates of the data are partitioned into n blocks each of length d in such a way that the entries in different blocks are independent but the entries from the same block may be dependent. In the second scenario – the random tensor model – the data is the homogeneous random tensor of order d, i.e. the coordinates of the data are all $\binom{n}{d}$ different products of d variables chosen from a set of n independent random variables. We show that Marchenko-Pastur law holds for the block-independent model as long as $d = o(n)$ and for the random tensor model as long as $d = o(n^{1/3})$. Our main technical tools are new concentration inequalities for quadratic forms in random variables with block-independent coordinates, and for random tensors.

1. Introduction

1.1. **Marchenko-Pastur law.** Consider a $p \times m$ random matrix X with independent entries that have zero mean and unit variance. The limiting distribution of eigenvalues $\lambda_i(W)$ of the sample covariance matrix $W = \frac{1}{m} XX^\mathsf{T}$ is determined by the celebrated Marchenko-Pastur law [35]. This result is valid in the regime where the dimensions of X increase to infinity but the aspect ratio converges to a constant, i.e. $p \to \infty$ and $p/m \to \lambda \in (0, \infty)$. Then, with probability 1, the empirical spectral distribution of the $p \times p$ matrix W converges weakly to a deterministic distribution that is now called the Marchenko-Pastur law with parameter λ. More specifically, if $\lambda \in (0, 1)$, then with probability 1 the following holds for each $x \in \mathbb{R}$:

$$F^W(x) := \frac{1}{p} \#\{1 \le i \le p : \ \lambda_i(W) \le x\} \to \int_\infty^x f_\lambda(t)\, dt$$

where f_λ is the Marchenko-Pastur density

$$f_\lambda(x) = \frac{1}{2\pi\lambda x} \sqrt{\big[(\lambda_+ - x)(x - \lambda_-)\big]_+}, \quad \text{with} \quad \lambda_\pm = (1 \pm \sqrt{\lambda})^2. \tag{1.1}$$

A similar result also holds for $\lambda > 1$, but in that case the the limiting distribution has an additional point mass of $1 - 1/\lambda$ at the origin. A straightforward proof of the Marchenko-Pastur law using the Stieltjes transform is given in Chapter 3 of [17]. More extensive expositions of the Marchenko-Pastur law with proofs using both the moment method and the Stieltjes transform are given in [10, Chapter 3] and [8]. Furthermore, [8] includes a review of many existing works prior to 1999.

Roman Vershynin is supported by USAF Grant FA9550-18-1-0031. Jennifer Bryson was partially supported by NSF Graduate Research Fellowship Program DGE-1321846. Hongkai Zhao was partially supported by NSF DMS-1622490 and DMS-1821010.

1.2. **Relaxing independence?** In many data sets it is natural to have independent columns, but not independent entries in the same column. For example, data collected from people, such as patient health information or personal movie ratings, will have independent columns since it is reasonable to assume each person's responses are independent of everyone else's responses. However, entries within a column are most likely not independent.

Several papers relaxing the independence within the columns already exist. Yin and Krishnaiah [52] required the columns X_k, to come from a spherically symmetric distribution; specifically, they require the distribution of X_k to be the same as that of PX_k where P is an orthogonal matrix. Aubrun [6] proved the result with X having independent columns distributed uniformly on the l_p^m ball. That result was generalized by Pajor and Pastur [39] who showed the independent columns can be distributed according to any isotropic log-concave measure. Götze and Tikhomirov in [21] and [22] replace the independence assumptions entirely with certain martinale-type conditions. In a similar manner, Adamczak [2] showed that Marchenko-Pastur law holds if the Euclidean norms of the rows and columns of X concentrate around their means and the expectation of each entry of X conditioned on all other entries equals zero. Bai and Zhou [11] gave a sufficient condition in terms of concentration of quadratic forms, and Yao [51] used their condition to allow a time series dependence structure in X. Yaskov [49] gave a short proof with a slightly weaker condition on the concentration of quadratic forms than Bai and Zhou's result. O'Rourke [38] considered a class of random matrices with dependent entries where even the columns are not necessarily independent, but are uncorrelated; although columns that are far enough apart must be independent. Lastly, the papers [15], [13], and [34] consider structured matrices such as block Toeplitz, Hankel, and Markov matrices.

In this paper, we study two new and natural models with relaxed independence requirement. In our first, block-independent model, we consider matrices with independent columns and each column is partitioned into blocks of the same size, and we only require the entries in different blocks to be independent.

Definition 1.1 (Block-independent model)**.** Consider a random vector $x \in \mathbb{R}^{nd}$. Assume that the entries of x can be partitioned into n blocks each of length d in such a way that the entries in different blocks are independent. (The entries from the same block may be dependent.) Then we say that x follows the block-independent model.

The block-independent data structures arise naturally in many situations. For example Netflix's movie recommendation data set contains ratings of movies by many people. A single person's movie ratings are likely to have a block structure coming from different movie genres, i.e. someone who dislikes documentary movies will have a block of poor ratings, etc. Another example of such a block structure is the stock market. The Marchenko-Pastur law assuming independence among all entries has been used as a comparison to the empirical spectral distribution of daily stock prices, see [29, 40]. However, we believe a block structure is more realistic, since for each day the performance of stocks in the same sector of the market are likely to be correlated and stocks in different sectors can be considered to be independent.

In the second model we study, the independent columns of a random matrix are formed by vectorized independent random tensors.

Definition 1.2 (Random tensor model)**.** Consider a random vector $x \in \mathbb{R}^n$ with independent entries. Let the random vector $\boldsymbol{x} \in \mathbb{R}^{\binom{n}{d}}$ be obtained by vectorizing the symmetric tensor $x^{\otimes d}$. Thus, the entries of $\boldsymbol{x}$ are indexed by d-element subsets $\mathbf{i} \subset [n]$ and are defined as products of the entries of x over $\mathbf{i}$:

$$\boldsymbol{x}_{\mathbf{i}} = \prod_{i \in \mathbf{i}} x_i = x_{i_1} x_{i_2} \ldots x_{i_d}, \quad \mathbf{i} = \{i_1, \ldots, i_d\}.$$

Then we say that $\boldsymbol{x}$ follows the random tensor model.

Although random tensors appear frequently in data science problems [5, 19, 26, 36, 48, 37, 41, 16, 14, 28, 24, 7, 46, 12, 54], a systematic theory of random tensors is still in its infancy.

1.3. **New results.** In this paper, we generalize Marchenko-Pastur law to the two models of random matrices described above. The following is our main result for the block-independence model (described in Definition 1.1 above). For simplicity, we first state it assuming that the entries from the same block are uncorrelated. We show how to remove this assumption in Section 1.6.

Theorem 1.3 (Marchenko-Pastur law for the block-independent model)**.** *Let $X = X^{(p)}$, $p = 1, 2, \ldots$, be a sequence of $p \times m$ random matrices, whose columns are independent and follow the block-independent model with $d = o(n)$ and uncorrelated entries within each block, and the aspect ratio p/m converge to a number $\lambda \in (0, \infty)$ as $p \to \infty$. Assume that all entries of the random matrix X have zero mean, unit variance, and uniformly bounded fourth moments. Then with probability 1 the empirical spectral distribution of the sample covariance matrix $W = \frac{1}{m} X X^\mathsf{T}$ converges weakly in distribution to the Marchenko-Pastur distribution with parameter λ.*

Remark 1.4 (Exchangeability removes any restrictions on block size). Suppose in the block-independent model we additionally have exchangeability in each block, meaning that

$$\mathbb{E}\, x_i x_j x_k x_l = \mathbb{E}\, x_q x_r x_s x_t$$

whenever the indices i, j, k, l, q, r, s, t are all in the same block. Then we will show that the Marchenko-Pastur law holds for such model *regardless of the size d of the blocks*, as long as the number of blocks $n \to \infty$.

Our second main result is the Marchenko-Pastur law for the random tensor model (described in Definition 1.2).

Theorem 1.5 (Marchenko-Pastur law for the random tensor model)**.** *Let $X = X^{(p)}$, $p = \binom{n}{d}, n = 1, 2, \ldots$, be a sequence of $p \times m$ random matrices, whose columns are independent and follow the random tensor model with $d = o(n^{1/3})$, and the aspect ratio p/m converge to a number $\lambda \in (0, \infty)$ as $p \to \infty$. Assume that the entries of the random vector x have zero mean, unit variance, and uniformly bounded fourth moments.*[1] *Then with probability 1 the empirical spectral distribution of the sample covariance matrix $W = \frac{1}{m} X X^\mathsf{T}$ converges weakly in distribution to the Marchenko-Pastur distribution with parameter λ.*

[1]Note that for the random tensor model, the fourth moment assumption only concerns the entries of the random *vector* x. The fourth moments of the entries of the random *tensor* $\boldsymbol{x}$, and thus of the entries of the random matrix X, can be very large. Indeed, if $\mathbb{E}\, x_i^4 = K$ for all i, then $\mathbb{E}\, \boldsymbol{x}_{\mathbf{i}}^4 = K^d$ by independence.

1.4. **Marchenko-Pastur law via concentration of quadratic forms.** Our approach to both main results is based on concentration of quadratic forms. Starting with the original proof of Marchenko-Pastur law [35] via Stieltjes transform, many arguments in random matrix theory (e.g. [39, 23]), make crucial use of concentration of quadratic forms. Specifically, at the core of the proof of Marchenko-Pastur law lies the bound

$$\mathrm{Var}(x^{\mathsf{T}}Ax) = o(p^2) \tag{1.2}$$

where $x \in \mathbb{R}^p$ is any column of the random matrix X and A is any deterministic $p \times p$ matrix with $\|A\| \le 1$. If the entires of x have uniformly bounded fourth moments, one always has

$$\mathrm{Var}(x^T Ax) = \mathbb{E}(x^T Ax)^2 \le \mathbb{E}\,\|x\|_2^4 = O(p^2).$$

Thus, the requirement (1.2) is just a little stronger than the trivial bound.

Suppose the columns of the random matrix X are independent, but the entries of each columns may be dependent. Then for Marchenko-Pastur to hold for X, it is sufficient (but not necessary) to verify the concentration inequality (1.2). The sufficiency is given in the following result; the absence of necessity is noted in [2, Section 2.1, Example 3].

Theorem 1.6 (Bai-Zhou [11])**.** *Let $X = X^{(p)}$, $p = 1, 2, \dots,$ be a sequence of mean zero $p \times m$ random matrices with independent columns. Assume the following as $p \to \infty$.*

1. *The aspect ratio p/m converges to a number $\lambda \in (0, \infty)$ as $p \to \infty$.*
2. *For each p, all columns X_k of $X^{(p)}$ have the same covariance matrix $\Sigma = \Sigma^{(p)} = \mathbb{E}\, X_k X_k^{\mathsf{T}}$. The spectral norm of the covariance matrix $\Sigma^{(p)}$ is uniformly bounded, and the empirical spectral distribution of $\Sigma^{(p)}$ converges to a deterministic distribution H.*
3. *For any deterministic $p \times p$ matrices $A = A^{(p)}$ with uniformly bounded spectral norm and for every column X_k, we have*
$$\max_k \mathrm{Var}(X_k^{\mathsf{T}} A X_k) = o(p^2).$$

Then, with probability 1 the empirical spectral distribution of the sample covariance matrix $W = \frac{1}{m} XX^{\mathsf{T}}$ converges weakly to a deterministic distribution whose Stieltjes transform satisfies

$$s(z) = \int_0^\infty \frac{1}{t(1-\lambda-\lambda z s) - z}\, dH(t), \quad z \in \mathbb{C}^+. \tag{1.3}$$

In the case where the entries of the columns are uncorrelated and have unit variance, we have $\Sigma = I$ and Theorem 1.6 yields that the limiting distribution is the original Marchenko-Pastur law (1.1).

1.5. **Concentration of quadratic forms: new results.** Theorem 1.6 reduces proving Marchenko-Pastur law for our new models to the concentration of a quadratic form $x^{\mathsf{T}}Ax$. If the random vector x has all independent entries, bounding the variance of this quadratic form is elementary. Moreover, in this case Hanson-Wright inequality (see e.g. [45, 42]) gives good probability tail bounds for the quadratic form.

But in our new models, the coordinates of the random vector are not independent. There seem to be no sufficiently powerful concentration inequalities available for such models. Known concentration inequalities for random chaoses [30, 32, 31, 3, 4, 20, 1] exhibit an unspecified (possibly exponential) dependence on the degree d, which is too bad for our purposes. An exception is the recent work [46] on concentration of random tensors with an optimal dependence on d. However, the results of [46] only apply for non-symmetric tensors and positive-semidefinite matrices A.

The following is a new concentration inequality for the block-independent model, which we will prove in Section 2.

Theorem 1.7 (Variance of quadratic forms for block-independent model)**.** *Let $x \in \mathbb{R}^{nd}$ be a random vector that follows the block-independent model. Then, for any fixed matrix $A \in \mathbb{R}^{nd\times nd}$, we have*

$$\mathrm{Var}(x^\mathsf{T} Ax) \le C\|A\|^2 Knd^3. \tag{1.4}$$

Here C is an absolute constant and K is the largest fourth moment of the entries of x. Moreover, if the entries of x within the same block are exchangeable, then the bound improves to

$$\mathrm{Var}(x^\mathsf{T} Ax) \le C\|A\|^2 Knd^2. \tag{1.5}$$

This result combined with Theorem 1.6 immediately establishes Marchenko-Pastur law for the block-independence model:

Proof of Theorem 1.3. In the block-independent model, the dimension is $p = nd$, and (1.4) becomes

$$\mathrm{Var}(x^\mathsf{T} Ax) \le C\|A\|^2 K\left(\frac{d}{n}\right)(nd)^2 = o(p^2)$$

whenever $\|A\| = O(1)$, $K = O(1)$, and $d = o(n)$. This justifies condition 3 of Theorem 1.6. Applying this theorem with $\Sigma = I$ we conclude Theorem 1.3. □

Remark 1.4 about exchangeable distributions can be derived similarly from (1.5).

Theorem 1.8 (Variance of quadratic forms for random tensor model)**.** *There exist positive absolute constants $C, c > 0$ such that the following holds. Let $\boldsymbol{x} \in \mathbb{R}^{\binom{n}{d}}$ be a random vector that follows the random tensor model. Then, for any fixed matrix $A \in \mathbb{R}^{\binom{n}{d}\times\binom{n}{d}}$, we have*

$$\mathrm{Var}(\boldsymbol{x}^\mathsf{T} A\boldsymbol{x}) \le C\|A\|^2 \binom{n}{d}^2 \left(\frac{K^{1/2}d}{n^{1/3}}\right)^{3/2},$$

as long as $K^{1/2}d/n^{1/3} < c$. Here K is the largest fourth moment of the entries of x.

This result combined with Theorem 1.6 immediately establishes Marchenko-Pastur law for the random tensor model:

Proof of Theorem 1.5. In the random tensor model, the dimension equals $p = \binom{n}{d}$ and (1.4) becomes

$$\mathrm{Var}(x^\mathsf{T} Ax) \le C\|A\|^2 K^{3/4}\left(\frac{d}{n^{1/3}}\right)^{3/2}\binom{n}{d}^2 = o(p^2)$$

whenever $\|A\| = O(1)$, $K = O(1)$, and $d = o(n^{1/3})$. This justifies condition 3 of Theorem 1.6. Applying this theorem with $\Sigma = I$ we conclude Theorem 1.5. □

1.6. **Allowing correlations in the block-independent model.** In Theorem 1.3, we assumed for simplicity that all entries have unit variance and the entires within each block are uncorrelated. We can remove both of these assumptions; the limiting spectral distribution will then be the anisotropic Marchenko-Pastur law (1.3).

To see this, suppose all columns of our random matrix $X = X^{(p)}$ have the same covariance matrix $\Sigma = \Sigma^{(p)}$. Assume that, as $p \to \infty$, we have $\|\Sigma^{(p)}\| = O(1)$ and the empirical spectral distribution[2] of $\Sigma^{(p)}$ converges to a deterministic distribution H.

Denoting as before by X_k the k-th column of X, we can represent it as $X_k = \Sigma^{1/2} x_k$ where x_k is some isotropic random vector, i.e. one whose entries are uncorrelated and have unit variance. Then

$$\mathrm{Var}(X_k^\mathsf{T} A X_k) = \mathrm{Var}\left(x_k^\mathsf{T}\, \Sigma^{1/2} A \Sigma^{1/2}\, x_k\right).$$

Applying Theorem 1.7 for $x = x_k$ and $\Sigma^{1/2} A \Sigma^{1/2}$ instead of A, we conclude that

$$\mathrm{Var}(X_k^\mathsf{T} A X_k) \le C \left\|\Sigma^{1/2} A \Sigma^{1/2}\right\|^2 K n d^3 \le C \|\Sigma\|^2 \|A\|^2 K \left(\frac{d}{n}\right)(nd)^2 = o(p^2)$$

whenever $\|\Sigma\| = O(1)$, $\|A\| = O(1)$, $K = O(1)$, and $d = o(n)$.

This justifies condition 3 of Theorem 1.6. Applying this theorem, we conclude that the limiting spectral distribution of $W = \frac{1}{m} X X^\mathsf{T}$ converges to the anisotropic Marchenko-Pastur distribution (1.3).

1.7. **Questions.** An obvious question is – how fast can $d = d(n)$ grow so that Marchenko-Pastur law holds? For the block-independent model, does Marchenko-Pastur law hold regardless of the size d of the blocks, as long as the number of blocks $n \to \infty$? We were only able to prove it under the exchangeability condition (Remark 1.4); otherwise Theorem 1.3 requires $d = o(n)$, which does not seem optimal. For the random tensor model, does Marchenko-Pastur law hold as long as $d = o(n)$? Our requirement $d = o(n^{1/3})$ in Theorem 1.5 is stronger and does not seem to be optimal.

What are the optimal bounds for concentration of quadratic forms in both of our models? The dependence on n and d we obtained in Theorem 1.7 and Theorem 1.8 does not seem to be optimal.

These two theorems give bounds on the variance. It would be interesting to go further and prove Hanson-Wright type concentration inequalities for the quadratic forms for both models.

2. Quadratic forms in block-independent random vectors: Proof of Theorem 1.7

Without loss of generality, we may assume that $\|A\| = 1$ by rescaling. Expanding $x^\mathsf{T} A x$ as a double sum of terms $A_{ij} x_i x_j$ and distinguishing the cases when $i = j$ or $i \ne j$, we have:

[2]Since the blocks are independent, the covariance matrix Σ is block-diagonal. If Σ_j is the covariance matrix of the block j, then the spectral norm of Σ is the maximal maximal spectral norm of Σ_j, and the empirical spectral distribution of Σ is the mixture of the empirical spectral distributions of all Σ_j.

$$\operatorname{Var}(x^{\mathsf T}Ax) = \mathbb{E}\left[|x^{\mathsf T}Ax - \operatorname{tr} A|^2\right] \le 2\,\mathbb{E}\left[\Big(\sum_{i=1}^{nd} A_{ii}(x_i^2-1)\Big)^2\right] + 2\,\mathbb{E}\left[\Big(\sum_{i\neq j} A_{ij}x_ix_j\Big)^2\right]$$
$$=: 2S_{\text{diag}} + 2S_{\text{off}}. \tag{2.1}$$

Here we used the inequality $(a+b)^2 \le 2a^2 + 2b^2$.

2.1. **Diagonal contribution.** Expanding the square, we can express the diagonal contribution as

$$S_{\text{diag}} = \sum_{i,k=1}^{nd} A_{ii}A_{kk}\,\mathbb{E}(x_i^2-1)(x_k^2-1).$$

Now if i and k are in different blocks, then by independence and unit variance, all such terms have expectation zero and do not contribute anything to S_{diag}. So, the only contribution comes from the terms where i and k are in the same block. In such a case, we have the bound

$$\mathbb{E}(x_i^2-1)(x_k^2-1) = \mathbb{E}\,x_i^2x_k^2 - 1 \le \max_\alpha \mathbb{E}\,x_\alpha^4 \le K.$$

Thus, the contribution of the terms for which j and k are in the *first* block is

$$\sum_{i,k=1}^{d} A_{ii}A_{kk}\,\mathbb{E}(x_i^2-1)(x_k^2-1) \le K\sum_{i,k=1}^{d} A_{ii}A_{kk} = Kd^2.$$

In the last step, we used that the magnitude of each entry of A is bounded by the spectral norm of A, which we assumed to be 1.

Clearly, the same result holds not only for the first block but for each of the n blocks. Summing up these bounds, we conclude that

$$S_{\text{diag}} \lesssim Knd^2. \tag{2.2}$$

2.2. **Off-diagonal contribution: setting up partitions.** Expanding the square, we can express the off-diagonal contribution in (2.1) as

$$S_{\text{off}} = \mathbb{E}\sum_{i\neq j}\sum_{k\neq l} A_{ij}A_{kl}x_ix_jx_kx_l, \tag{2.3}$$

we can break this into cases considering partitions of 4, representing the powers on the entries of x. For example, the partition $(2,1,1)$ represents the indices (i,j,k,l) for which $i\neq j$, $k\neq l$ and such that exactly two among the four indices are the same. This comprises the terms of the form $A_{ij}A_{il}x_i^2x_jx_l$, $A_{ij}A_{ki}x_i^2x_jx_k$, $A_{ij}A_{jl}x_ix_j^2x_l$, and $A_{ij}A_{kj}x_ix_j^2x_k$.

Notice that in all partitions of 4 we have to consider, i.e. those for which $i\neq j$ and $k\neq l$, none of the powers can be greater than 2. (Indeed, in the partition $(3,1)$ three indices among i,j,k,l are the same, which violates either the constraint $i\neq j$ or the constraint $k\neq l$.) This leaves us with partitions $(2,2)$, $(2,1,1)$ and $(1,1,1,1)$, which we shall consider one by one.

2.3. **Partition** $(2,2)$**.** The terms corresponding to this partition have the form $A_{ij}^2 x_i^2 x_j^2$ and $A_{ij}A_{ji}x_i^2x_j^2$. Notice that

$$\mathbb{E}[x_i^2 x_j^2] \le \big(\mathbb{E}[x_i^4]\big)^{1/2}\big(\mathbb{E}[x_j^4]\big)^{1/2} \le K \tag{2.4}$$

Thus, we can bound the net contribution of the terms of the form $A_{ij}^2 x_i^2 x_j^2$ as follows:

$$\mathbb{E}\sum_{\substack{i,j=1\\ i\ne j}}^{nd} A_{ij}^2 x_i^2 x_j^2 \le K \sum_{i,j=1}^{nd} A_{ij}^2 \le Knd$$

where in the last step we used the bound

$$\sum_{i,j=1}^{nd} A_{ij}^2 = \|A\|_F^2 \le nd\|A\|^2 = nd. \tag{2.5}$$

Similarly, we can bound the net contribution of the terms of the form $A_{ij}A_{ji}x_i^2x_j^2$:

$$\mathbb{E}\sum_{\substack{i,j=1\\ i\ne j}}^{nd} A_{ij}A_{ji}x_i^2x_j^2 \le K\sum_{i,j=1}^{nd}|A_{ij}||A_{ji}| \le K\Big(\sum_{i,j=1}^{nd}A_{ij}^2\Big)^{1/2}\Big(\sum_{i,j=1}^{nd}A_{ji}^2\Big)^{1/2} \le Knd$$

where we used the moment bound (2.4), Cauchy-Schwarz inequality, and (2.5).

Concluding, the net contribution to (2.3) of the terms comprising the partition $(2,2)$ is $\lesssim Knd$.

2.4. **Partition** $(1,1,1,1)$**.** The terms corresponding to this partition have the form $A_{ij}A_{kl}x_ix_jx_kx_l$ where all indices i,j,k,l are distinct.

We claim that the expectation of such a term is zero unless all four indices i,j,k,l come from *the same block*. Indeed, suppose one of these indices – let's say i – comes from its own block, and none of the other three indices i,j,k are in the same block as i. Then $\mathbb{E}\, x_ix_jx_kx_l = 0$ since in this case x_i is independent of x_j, x_k and x_l and has zero mean. The remaining situation is where a pair of the indices – let's say i,j – are in one block, and the other pair k,l is in a different block. Then again $\mathbb{E}\, x_ix_jx_kx_l = \mathbb{E}[x_ix_j]\,\mathbb{E}[x_kx_l] = 0$ since x_ix_j is independent of x_kx_l, and moreover x_i and x_j are uncorrelated. This proves the claim.

So, let us assume that all indices i,j,k,l are in the same block. In such a case, we have the bound

$$\mathbb{E}\, x_ix_jx_kx_l \le \max_\alpha \mathbb{E}[x_\alpha^4] \le K. \tag{2.6}$$

Thus, the contribution of the terms for which i,j,k,l are in the *first* block can be bounded by

$$\mathbb{E}\sum_{\substack{i,j,k,l=1\\ i,j,k,l \text{ distinct}}}^{d} A_{ij}A_{kl}x_ix_jx_kx_l \le K\sum_{i,j,k,l=1}^{d}|A_{ij}A_{kl}| = K\Big(\sum_{i,j=1}^{d}|A_{ij}|\Big)^2 \le Kd^2\sum_{i,j=1}^{d}A_{ij}^2 \le Kd^3. \tag{2.7}$$

In the last step, we used that top-left $d\times d$ minor of A, which we denote by $A_{d\times d}$, satisfies

$$\sum_{i,j=1}^{d} A_{ij}^2 = \|A_{d\times d}\|_F^2 \le d\|A_{d\times d}\|^2 \le d\|A\|^2 = d.$$

Clearly, a bound similar to (2.7) holds not only for the first block but for each of the n blocks. Summing up these bounds, we conclude that the net contribution to (2.3) of the terms corresponding to the partition $(1,1,1,1)$ is $\lesssim Knd^3$.

2.5. **Partition** $(2,1,1)$**.** This partition comprises the terms of the form $A_{ij}A_{il}x_i^2x_jx_l$, $A_{ij}A_{ki}x_i^2x_jx_k$, $A_{ij}A_{jl}x_ix_j^2x_l$, and $A_{ij}A_{kj}x_ix_j^2x_k$. Just like in the previous case where we studied the partition $(1,1,1,1)$, we can argue that the only nonzero contribution comes from the terms where all three indices are *in the same block*.

Let us consider the terms of the form $A_{ij}A_{il}x_i^2x_jx_l$ first. In such a case, we have the bound

$$\mathbb{E}\, x_i^2x_jx_l \le \max_\alpha \mathbb{E}[x_\alpha^4] \le K.$$

Thus, the contribution of the terms for which i,j,l are in the *first* block can be bounded by

$$\mathbb{E} \sum_{\substack{i,j,l=1 \\ i,j,l \text{ distinct}}}^{d} A_{ij}A_{il}x_i^2x_jx_l \le K \sum_{i,j,l=1}^{d} |A_{ij}A_{il}| = K \sum_{i=1}^{d} \left(\sum_{j=1}^{d} |A_{ij}| \right)^2 \le Kd^2.$$

In the last step, we used that top-left $d \times d$ minor of A, which we denote by $A_{d\times d}$, satisfies

$$\sum_{j=1}^{d} |A_{ij}| \le \sqrt{d} \left(\sum_{j=1}^{d} A_{ij}^2 \right)^{1/2} \le \sqrt{d}\|A_{d\times d}\| = \sqrt{d} \quad \text{for every } i.$$

A similar result holds not only for the first block but for each of the n blocks. Summing up these bounds, we conclude that the net contribution to (2.3) of the terms of the form $A_{ij}A_{il}x_i^2x_jx_l$ is $\lesssim Knd^2$. Finally, we can repeat the above argument for the terms of the other three types, $A_{ij}A_{ki}x_i^2x_jx_k$, $A_{ij}A_{jl}x_ix_j^2x_l$, and $A_{ij}A_{kj}x_ix_j^2x_k$, and thus conclude that the net contribution to (2.3) of the terms corresponding to the partition $(2,1,1)$ is $\lesssim Knd^2$.

Ultimately, adding the contributions of all partitions – $(2,2)$, $(1,1,1,1)$, and $(2,1,1)$, we see that the total off-diagonal contribution (2.3) is bounded by

$$S_{\text{off}} = O(Knd) + O(Knd^3) + O(Knd^2) = O(Knd^3).$$

Combining this with the bound (2.2) on the diagonal contribution and plugging into (2.1), we conclude that

$$\operatorname{Var}(x^\mathsf{T} Ax) = O(Knd^2) + O(Knd^3) = O(Knd^3).$$

This proves the first conclusion of Theorem 1.7. □

2.6. **Exchangeable distributions.** Here we prove the second part of the Theorem 1.7. Namely, we assume that the entries of x within the same block are exchangeable, and we seek to improve our previous bound on $\operatorname{Var}(x^\mathsf{T} Ax)$ from $O(Knd^3)$ to $O(Knd^2)$. A quick look at the previous section reveals that the only part that needs to be strengthened is the partition $(1,1,1,1)$, where our bound was $O(Knd^3)$; it suffices to improve it to $O(Knd^2)$.

So let us focus on the partition $(1,1,1,1)$. Thus, we will have to bound the sum of the terms $A_{ij}A_{kl}x_ix_jx_kx_l$ over quadruples i,j,k,l of all distinct indices. As we argued in the beginning of Section 2.4, we can assume without generality that for each term, the indices i,j,k,l are in *the same block* (otherwise the expectation of such a term would be zero). Focusing on the *first* block for now, we are seeking to bound the quantity

(2.8)
$$\Big|\mathbb{E}\sum_{\substack{i,j,k,l=1\\ i,j,k,l \text{ distinct}}}^{d} A_{ij}A_{kl}x_ix_jx_kx_l\Big| = \Big|\mathbb{E}[x_1x_2x_3x_4]\sum_{\substack{i,j,k,l=1\\ i,j,k,l \text{ distinct}}}^{d} A_{ij}A_{kl}\Big| \le K\Big|\sum_{\substack{i,j,k,l=1\\ i,j,k,l \text{ distinct}}}^{d} A_{ij}A_{kl}\Big|,$$

where the equality follows from the exchangeability assumption, and the inequality follows from (2.6). We reduced the problem to bounding the sum of $A_{ij}A_{kl}$ over quadruples i, j, k, l of all distinct indices in $\{1, \dots, d\}$. The following lemma provides an adequate bound.

Lemma 2.1. *Any real $d \times d$ matrix B satisfies*
$$\Big|\sum_{\substack{i,j,k,l=1\\ i,j,k,l \text{ distinct}}}^{d} B_{ij}B_{kl}\Big| \le 10d^2\|B\|^2.$$

Proof. Without loss of generality, we can assume that $\|B\| = 1$. Denote by $\mathbf{1} \in \mathbb{R}^d$ the vector with all 1 coordinates and note that
$$\Big|\sum_{i,j=1}^{d} B_{ij}\Big| = |\mathbf{1}^\mathsf{T} B\mathbf{1}| \le \|\mathbf{1}\|_2^2\|B\| = d \quad\text{and}\quad \Big|\sum_{i=1}^{d} B_{ii}\Big| = |\operatorname{tr} B| \le d\|B\| = d.$$

Subtracting one sum from the other and using triangle inequality, we get
$$\Big|\sum_{\substack{i,j=1\\ i\ne j}}^{d} B_{ij}\Big| \le 2d.$$

This yields

$$\Big|\sum_{\substack{i,j,k,l=1\\ i\ne j,\ k\ne l}}^{d} B_{ij}B_{kl}\Big| = \Big(\sum_{\substack{i,j=1\\ i\ne j}}^{d} B_{ij}\Big)^2 \le 4d^2. \tag{2.9}$$

Thus, instead of the requirement that all four indices i, j, k, l be distinct, we were able to handle the weaker but simpler requirement that $i \ne j$ and $k \ne l$. This weaker requirement produces the sum over more terms. It remains to control the sum over the difference set E, i.e. over the set of quadruples of indices i, j, k, l not all of which are distinct but for which $i \ne j$ and $k \ne l$.

This set E can be expressed as the union of the following four sets:
$$E_1 := \big\{(i,j,k,l) :\ j \ne i = k \ne l\big\}, \qquad E_2 := \big\{(i,j,k,l) :\ j \ne i = l \ne k\big\},$$
$$E_3 := \big\{(i,j,k,l) :\ i \ne j = k \ne l\big\}, \qquad E_4 := \big\{(i,j,k,l) :\ i \ne j = l \ne k\big\}.$$

By the inclusion-exclusion principle, the sum of any terms w_{ijkl} over E can be expressed as

$$\sum_E w_{ijkl} = \sum_{E_1} w_{ijkl} + \sum_{E_2} w_{ijkl} + \sum_{E_3} w_{ijkl} + \sum_{E_4} w_{ijkl} - \sum_{E_1\cap E_4} w_{ijkl} - \sum_{E_2\cap E_3} w_{ijkl}, \tag{2.10}$$

since only two of all pairwise intersections are nonempty, namely $E_1 \cap E_4$ and $E_2 \cap E_3$, and all three-wise and four-wise intersections are empty.

The sum over E_α can be bounded by the same argument as in Section 2.5, which we repeat here for the reader's convenience. For example, let us bound the sum over E_1:

$$\Big| \sum_{\substack{i,j,l=1 \\ i \neq j}}^{d} B_{ij} B_{il} \Big| \le \sum_{i,j,l=1}^{d} |B_{ij}||B_{il}| = \sum_{i=1}^{d} \Big(\sum_{j=1}^{d} |B_{ij}| \Big)^2 \le d^2.$$

In the last step, we used that

$$\sum_{j=1}^{d} |B_{ij}| \le \sqrt{d} \Big(\sum_{j=1}^{d} B_{ij}^2 \Big)^{1/2} \le \sqrt{d} \|B\| = \sqrt{d} \quad \text{for every } i.$$

Similarly one can bound the sums over E_2, E_3, and E_4.

The sum over the pairwise intersections is simpler to bound. For example, let us bound the sum over $E_2 \cap E_3$:

$$\Big| \sum_{\substack{i,j=1 \\ i \neq j}}^{d} B_{ij} B_{ji} \Big| \le \sum_{i,j=1}^{d} |B_{ij}||B_{ji}| \le d^2.$$

In the last step, we used that the magnitude of each entry of B is bounded by the spectral norm of B, which we assumed to be 1. The sum over $E_1 \cap E_4$ can be handled similarly.

Thus, each of the six sums on the right in the inclusion-exclusion formula (2.10) is bounded by d^2. Hence the sum over E is bounded by $6d^2$. Thus, the difference of the sum over all-distinct i, j, k, l and over the larger set where $i \neq j$, $k \neq l$ can be at most $6d^2$. Using (2.9), this implies that sum over all-distinct i, j, k, l is at most $4d^2 + 6d^2 \le 10d^2$. The lemma is proved. □

Apply Lemma 2.1 for the top-left $d \times d$ minor of A, and substitute into (2.8). We obtain

$$\Big| \mathbb{E} \sum_{\substack{i,j,k,l=1 \\ i,j,k,l \text{ distinct}}}^{d} A_{ij} A_{kl} x_i x_j x_k x_l \Big| \le 10 K d^2.$$

A similar bound clearly holds not only when the indices i, j, k, l are in the first block, but also for each of the n blocks. Summing up these bounds, we conclude that the net contribution of the terms corresponding to the partition $(1, 1, 1, 1)$ is $\lesssim Knd^2$. As we explained in the beginning of this section, this completes the proof of the second part of Theorem 1.7. □

3. Quadratic forms in random tensors: Proof of Theorem 1.8

The first steps of the proof are the same as for Theorem 1.7. Without loss of generality, we may assume that $\|A\| = 1$ by rescaling. Expanding $\boldsymbol{x}^\mathsf{T} A \boldsymbol{x}$ as a double sum of terms $A_{\mathbf{ij}} \boldsymbol{x}_{\mathbf{i}} \boldsymbol{x}_{\mathbf{j}}$, and distinguishing the diagonal terms ($\mathbf{i} = \mathbf{j}$) and the off-diagonal terms ($\mathbf{i} \neq \mathbf{j}$), we have:

$$\begin{aligned} \operatorname{Var}(\boldsymbol{x}^\mathsf{T} A \boldsymbol{x}) &\le 2\,\mathbb{E}\Big[\Big(\sum_{\mathbf{i}} A_{\mathbf{ii}} (\boldsymbol{x}_{\mathbf{i}}^2 - 1) \Big)^2 \Big] + 2\,\mathbb{E}\Big[\Big(\sum_{\mathbf{i} \neq \mathbf{j}} A_{\mathbf{ij}} \boldsymbol{x}_{\mathbf{i}} \boldsymbol{x}_{\mathbf{j}} \Big)^2 \Big] \\ &=: 2S_{\text{diag}} + 2S_{\text{off}}. \end{aligned} \tag{3.1}$$

3.1. **Diagonal contribution.** Expanding the square, we can express the diagonal contribution as

$$S_{\text{diag}} = \sum_{\mathbf{i},\mathbf{k}} A_{\mathbf{ii}} A_{\mathbf{kk}}\, \mathbb{E}(\boldsymbol{x}_{\mathbf{i}}^2 - 1)(\boldsymbol{x}_{\mathbf{k}}^2 - 1). \tag{3.2}$$

Both meta-indices $\mathbf{i}$ and $\mathbf{k}$ range in all $\binom{n}{d}$ subsets of $[n]$ of cardinality d. Let v denote the overlap between these two subsets, i.e.

$$v := |\mathbf{i} \cap \mathbf{k}|.$$

If $v = 0$, the subsets are disjoint, the random variables $\boldsymbol{x}_{\mathbf{i}}^2 - 1$ and $\boldsymbol{x}_{\mathbf{k}}^2 - 1$ are independent and have mean zero, and thus

$$\mathbb{E}(\boldsymbol{x}_{\mathbf{i}}^2 - 1)(\boldsymbol{x}_{\mathbf{k}}^2 - 1) = 0.$$

Such terms do not contribute anything to the sum in (3.2).

If $v \geq 1$, the monomial $\boldsymbol{x}_{\mathbf{i}}^2 \boldsymbol{x}_{\mathbf{k}}^2$ consists of v terms raised to the fourth power (coming from the indices that are both in $\mathbf{i}$ and $\mathbf{k}$) and $2(d-v)$ terms raised to the second power (coming from the symmetric difference of $\mathbf{i}$ and $\mathbf{k}$). Thus,

$$\left| \mathbb{E}(\boldsymbol{x}_{\mathbf{i}}^2 - 1)(\boldsymbol{x}_{\mathbf{k}}^2 - 1) \right| \leq \mathbb{E}\, \boldsymbol{x}_{\mathbf{i}}^2 \boldsymbol{x}_{\mathbf{k}}^2 \leq \max_{\alpha} \left(\mathbb{E}\, x_\alpha^4 \right)^v \cdot \max_{\beta} \left(\mathbb{E}\, x_\beta^2 \right)^{2(d-v)} \leq K^v,$$

where we used the unit variance assumption.

There are $\binom{n}{d}$ ways to choose $\mathbf{i}$. Once we fix $\mathbf{i}$ and $v \in \{1, \ldots, d\}$, there are $\binom{d}{v}\binom{n-d}{d-v}$ ways to choose $\mathbf{k}$, since v indices must come from $\mathbf{i}$ and the remaining $d - v$ indices must come from $[n] \setminus \mathbf{i}$. Therefore,

$$S_{\text{diag}} \leq \binom{n}{d} \sum_{v=1}^{d} \binom{d}{v} \binom{n-d}{d-v} K^v. \tag{3.3}$$

To bound this sum, we can assume without loss of generality that K is a positive integer. Then the following elementary inequality holds:

$$\binom{d}{v} K^v \leq \binom{Kd}{v},$$

and it can be quickly checked by writing the binomial coefficients in terms of factorials. Now, if we were summing v from zero as opposed from 1 in (3.3), we can use Vandermonde's identity and get

$$\sum_{v=0}^{d} \binom{d}{v} \binom{n-d}{d-v} K^v \leq \sum_{v=0}^{d} \binom{Kd}{v} \binom{n-d}{d-v} = \binom{n-d+Kd}{d}.$$

Subtracting the zeroth term, we obtain

$$\sum_{v=1}^{d} \binom{d}{v} \binom{n-d}{d-v} K^v \leq \binom{n-d+Kd}{d} - \binom{n-d}{d}.$$

Now use a stability property of binomial coefficients (Lemma 3.7), which tells us that

$$\binom{n-d+Kd}{d} - \binom{n-d}{d} \leq \delta \binom{n-d}{d} \quad \text{where } \delta := \frac{2Kd^2}{n-2d+1},$$

as long as $\delta \le 1/2$. According to our assumptions on the degree d, we do have $\delta \le 1/2$ when n is sufficiently large.

Summarizing, we have shown that

$$S_{\mathrm{diag}} \le \binom{n}{d} \cdot \delta \binom{n-d}{d} \lesssim \binom{n}{d}^2 \cdot \frac{Kd^2}{n}. \tag{3.4}$$

3.2. **Off-diagonal contribution: the cross moments.** Expanding the square, we can express the off-diagonal contribution in (3.1) as

$$S_{\mathrm{off}} = \sum_{\mathbf{i} \ne \mathbf{j}} \sum_{\mathbf{k} \ne \mathbf{l}} A_{\mathbf{ij}} A_{\mathbf{kl}} \, \mathbb{E}\, \boldsymbol{x}_{\mathbf{i}} \boldsymbol{x}_{\mathbf{j}} \boldsymbol{x}_{\mathbf{k}} \boldsymbol{x}_{\mathbf{l}}. \tag{3.5}$$

Let us first bound the expectation of

$$\boldsymbol{x}_{\mathbf{i}} \boldsymbol{x}_{\mathbf{j}} \boldsymbol{x}_{\mathbf{k}} \boldsymbol{x}_{\mathbf{l}} = \prod_{i \in \mathbf{i}} x_i \prod_{i \in \mathbf{i}} x_j \prod_{k \in \mathbf{k}} x_k \prod_{l \in \mathbf{l}} x_l.$$

Without loss of generality, we can assume that this monomial of degree $4d$ has no linear factors, i.e. each of the factors x_α of this monomial has degree at least 2, otherwise the expectation of the monomial is zero. Rearranging the factors, we can express the monomial as

$$\boldsymbol{x}_{\mathbf{i}} \boldsymbol{x}_{\mathbf{j}} \boldsymbol{x}_{\mathbf{k}} \boldsymbol{x}_{\mathbf{l}} = \prod_{\alpha \in \Lambda_2} x_\alpha^2 \prod_{\beta \in \Lambda_3} x_\beta^3 \prod_{\gamma \in \Lambda_4} x_\gamma^4 \tag{3.6}$$

for some disjoint sets $\Lambda_2, \Lambda_3, \Lambda_4 \subset [n]$. Thus, Λ_2 consists of the indices that are covered by exactly two of the sets $\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l}$, and similarly for Λ_3 and Λ_4. Since each of the four sets $\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l}$ contains d indices, counting the indices with multiplicities gives

$$4d = 2|\Lambda_2| + 3|\Lambda_3| + 4|\Lambda_4|. \tag{3.7}$$

Since each index is covered at least by two of the four sets $\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l}$, the cardinality of the set

$$\mathbf{i} \cup \mathbf{j} \cup \mathbf{k} \cup \mathbf{l} = \Lambda_2 \sqcup \Lambda_3 \sqcup \Lambda_4 \tag{3.8}$$

is at most $4d/2 = 2d$. Let $w \ge 0$ be the "defect" defined by

$$\left| \mathbf{i} \cup \mathbf{j} \cup \mathbf{k} \cup \mathbf{l} \right| = 2d - w. \tag{3.9}$$

Thus, w would be zero if every index is covered by exactly two sets, and w would be positive if there are triple or quadruple covered indices. From (3.8) and (3.9) we see that

$$2d - w = |\Lambda_2| + |\Lambda_3| + |\Lambda_4|.$$

Multiplying both sides of this equation by 2 and subtracting from (3.7), we get

$$2w = |\Lambda_3| + 2|\Lambda_4|, \tag{3.10}$$

a relation that will be useful in a moment.

Take expectation on both sides of (3.6). Using independence and the assumptions that $\mathbb{E}\, x_\alpha^2 = 1$ and $\mathbb{E}\, \boldsymbol{x}_\alpha^4 \le K$ for each α, we get

$$\mathbb{E}\, |\boldsymbol{x}_{\mathbf{i}} \boldsymbol{x}_{\mathbf{j}} \boldsymbol{x}_{\mathbf{k}} \boldsymbol{x}_{\mathbf{l}}| = \prod_{\beta \in \Lambda_3} \mathbb{E}\, |x_\beta|^3 \cdot \prod_{\gamma \in \Lambda_4} \mathbb{E}\, x_\gamma^4 = \prod_{\beta \in \Lambda_3} \left(\mathbb{E}\, |x_\beta|^4 \right)^{3/4} \cdot \prod_{\gamma \in \Lambda_4} \mathbb{E}\, x_\gamma^4 \le K^{\frac{3}{4}|\Lambda_3| + |\Lambda_4|}.$$

Due to (3.10),

$$\frac{3}{4}|\Lambda_3| + |\Lambda_4| = \frac{3}{2}w - \frac{1}{2}|\Lambda_4| \le \frac{3}{2}w.$$

Thus we have shown that

$$\mathbb{E}\,|\boldsymbol{x}_{\mathbf{i}}\boldsymbol{x}_{\mathbf{j}}\boldsymbol{x}_{\mathbf{k}}\boldsymbol{x}_{\mathbf{l}}| \le K^{3w/2}.$$

3.3. **Sizes of intersections of meta-indices.** Due to the last step, the off-diagonal contribution (3.5) can be bounded as follows:

$$S_{\text{off}} \le \sum_{\mathbf{i}\ne\mathbf{j}} \sum_{\mathbf{k}\ne\mathbf{l}} |A_{\mathbf{ij}}||A_{\mathbf{kl}}|K^{3w/2}, \tag{3.11}$$

where the sum only includes the sets $\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l}$ that provide at least a *double cover*, i.e. such that every index from $\mathbf{i}\cup\mathbf{j}\cup\mathbf{k}\cup\mathbf{l}$ must belong to at least two of these four sets. We quantified this property by the *defect* $w \ge 0$, which we defined by

$$|\mathbf{i}\cup\mathbf{j}\cup\mathbf{k}\cup\mathbf{l}| = |\mathbf{i}\cup\mathbf{j}\cup\mathbf{k}| = 2d - w.$$

In preparation to bounding the double sum in (3.11), let us consider

$$|\mathbf{i}\cap\mathbf{j}| =: v, \quad |\mathbf{i}\cap\mathbf{j}\cap\mathbf{k}| =: r,$$

and observe a few useful bounds involving w, v, and r.

Lemma 3.1. *We have $w \le v \le d-1$.*

Proof. By definition, $v = |\mathbf{i}\cap\mathbf{j}| \le |\mathbf{i}| = d$. Moreover, v may not equal d, for this would mean that $\mathbf{i} = \mathbf{j}$, a possibility that is excluded in the double sum (3.11). This means that $v \le d-1$. Next, we have

$$|\mathbf{i}\cup\mathbf{j}| = |\mathbf{i}| + |\mathbf{j}| - |\mathbf{i}\cap\mathbf{j}| = 2d - v. \tag{3.12}$$

On the other hand, $|\mathbf{i}\cup\mathbf{j}| \le |\mathbf{i}\cup\mathbf{j}\cup\mathbf{k}| = 2d - w$. Combining these two facts yields $w \le v$. □

Lemma 3.2. *We have $r \le v$ and $r \le 2w$.*

Proof. The first statement follows from definition. To prove the second statement, recall that the sets $\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l}$ form at least a double cover of $\mathbf{i}\cup\mathbf{j}\cup\mathbf{k}\cup\mathbf{l}$ and at least a triple cover of $\mathbf{i}\cap\mathbf{j}\cap\mathbf{k}$ (trivially). Since each of the four sets has d indices, counting the indices with multiplicities gives

$$4d \ge 2|\mathbf{i}\cup\mathbf{j}\cup\mathbf{k}\cup\mathbf{l}| + |\mathbf{i}\cap\mathbf{j}\cap\mathbf{k}| = 2(2d-w) + r$$

by the definition of w and r. This yields $r \le 2w$. □

Lemma 3.3. *We have $r \le d - v + w$.*

Proof. The sets $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ obviously form at least a double cover of $\mathbf{i}\cap\mathbf{j}$ and a triple cover of $\mathbf{i}\cap\mathbf{j}\cap\mathbf{k}$. Since each of the three sets has d indices, counting the indices with multiplicities gives

$$3d \ge |\mathbf{i}\cup\mathbf{j}\cup\mathbf{k}| + |\mathbf{i}\cap\mathbf{j}| + |\mathbf{i}\cap\mathbf{j}\cap\mathbf{k}| = (2d-w) + v + r$$

by definition of w, v and r. Rearranging the terms completes the proof. □

3.4. **Number of choices of meta-indices.** Let us fix w, v, and r, and estimate the number of possible choices for the sets $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$, $\mathbf{l}$ that conform to these w, v, and r. This would help us determining the number of terms in the double sum (3.11). Thus, we would like to know how many ways are there to choose four d-element sets $\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l} \subset [n]$ that provide at least a double cover of $\mathbf{i} \cup \mathbf{j} \cup \mathbf{k} \cup \mathbf{l}$, and so that

$$|\mathbf{i} \cup \mathbf{j} \cup \mathbf{k}| = |\mathbf{i} \cup \mathbf{j} \cup \mathbf{k} \cup \mathbf{l}| = 2d - w, \quad |\mathbf{i} \cap \mathbf{j}| = v, \quad \text{and} \quad |\mathbf{i} \cap \mathbf{j} \cap \mathbf{k}| = r. \tag{3.13}$$

Choosing $\mathbf{i}$. This is easy: there are $\binom{n}{d}$ ways to choose the d-element subset $\mathbf{i}$ from $[n]$.

Choosing $\mathbf{j}$. Recall that we need to obey $|\mathbf{i} \cap \mathbf{j}| = v$. Thus, for a fixed $\mathbf{i}$, we have $\binom{d}{v}\binom{n-d}{d-v}$ choices for $\mathbf{j}$, which is seen by first picking the v overlapping indices from $\mathbf{i}$ and then the remaining $d - v$ indices from $\mathbf{i}^c$.

Choosing $\mathbf{k}$. Let us fix $\mathbf{i}$ and $\mathbf{j}$. The set of all available indices $[n]$, from which the indices of $\mathbf{k}$ can be chosen, can be partitioned into the three disjoint sets:

$$[n] = (\mathbf{i} \cap \mathbf{j}) \sqcup (\mathbf{i} \cup \mathbf{j})^c \sqcup (\mathbf{i} \triangle \mathbf{j}). \tag{3.14}$$

Let us see how many indices for $\mathbf{k}$ should come from each of these three sets.

As we see from (3.13), the v-element set $\mathbf{i} \cap \mathbf{j}$ must contain exactly r indices of k, and these can be selected in $\binom{v}{r}$ ways.

Next, we know from (3.12) that $|(\mathbf{i} \cup \mathbf{j})^c| = n - (2d - v)$, and

$$|(\mathbf{i} \cup \mathbf{j})^c \cap \mathbf{k}| = |\mathbf{i} \cup \mathbf{j} \cup \mathbf{k}| - |\mathbf{i} \cup \mathbf{j}| = (2d - w) - (2d - v) = v - w, \tag{3.15}$$

where we used (3.13) and (3.12). So, the set $(\mathbf{i} \cup \mathbf{j})^c$ must contain exactly $v - w$ indices of $\mathbf{k}$, and these can be selected in $\binom{n-(2d-v)}{v-w}$ ways.[3]

Finally, by (3.12) and (3.13) we have

$$|\mathbf{i} \triangle \mathbf{j}| = |\mathbf{i} \cup \mathbf{j}| - |\mathbf{i} \cap \mathbf{j}| = (2d - v) - v = 2(d - v). \tag{3.16}$$

We already allocated $r + (v - w)$ indices of k to the first two sets on the right-hand side of (3.14). Thus, the number of indices for $\mathbf{k}$ that come from the third set, $\mathbf{i} \triangle \mathbf{j}$, must be

$$|(\mathbf{i} \triangle \mathbf{j}) \cap \mathbf{k}| = d - r - (v - w). \tag{3.17}$$

These indices can be selected in $\binom{2(d-v)}{d-r-(v-w)}$ ways.[4]

Summarizing, for fixed $\mathbf{i}$ and $\mathbf{j}$, we have $\binom{v}{r}\binom{n-(2d-v)}{v-w}\binom{2(d-v)}{d-r-(v-w)}$ choices for $\mathbf{k}$.

Choosing $\mathbf{l}$. Fix $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$. Recall that the sets $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$, $\mathbf{l}$ must form at least a double cover of $\mathbf{i} \cup \mathbf{j} \cup \mathbf{k} \cup \mathbf{l}$. This has two consequences. First, we must have

$$\mathbf{l} \subset \mathbf{i} \cup \mathbf{j} \cup \mathbf{k} \tag{3.18}$$

to avoid any single-covered indices in $\mathbf{l}$. Second, $\mathbf{l}$ must contain all the *single indices*, i.e. those that belong to exactly one of the sets $\mathbf{i}$, $\mathbf{j}$, or $\mathbf{k}$. The set of single indices, denoted $\mathbf{s}$, can be represented as

$$\mathbf{s} = (\mathbf{i}^c \cap \mathbf{j}^c \cap \mathbf{k}) \sqcup \big[(\mathbf{i} \cap \mathbf{j}^c \cap \mathbf{k}^c) \sqcup (\mathbf{i}^c \cap \mathbf{j} \cap \mathbf{k}^c)\big] = \big[(\mathbf{i} \cup \mathbf{j})^c \cap \mathbf{k}\big] \sqcup \big[(\mathbf{i} \triangle \mathbf{j}) \cap \mathbf{k}^c\big].$$

[3]Since the cardinality of any set is nonnegative, equation (3.15) provides an alternative proof of the bound $w \le v$ in Lemma 3.1.

[4]Since the cardinality of any set is nonnegative, equation (3.17) provides an alternative proof of Lemma 3.3.

At this stage, the sets $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are all fixed, and so is $\mathbf{s}$.

To compute the cardinality of $\mathbf{s}$, recall from (3.15) that $|(\mathbf{i} \cup \mathbf{j})^c \cap \mathbf{k}| = v - w$. Furthermore, using (3.16) and (3.17), we see that

$$|(\mathbf{i} \triangle \mathbf{j}) \cap \mathbf{k}^c| = |(\mathbf{i} \triangle \mathbf{j})| - |(\mathbf{i} \triangle \mathbf{j}) \cap \mathbf{k}| = 2(d-v) - (d - r - (v-w)) = d - w - v + r.$$

Thus, the number of single indices is

$$|\mathbf{s}| = (v - w) + (d - w - v + r) = d - 2w + r.$$

Since $\mathbf{l}$ must contain the set $\mathbf{s}$ of single indices, which is fixed, the only freedom in choosing $\mathbf{l}$ comes from selecting non-single indices. There are $d - (d - 2w + r) = 2w - r$ of them,[5] and they must come from the set $(\mathbf{i} \cup \mathbf{j} \cup \mathbf{k}) \setminus \mathbf{s}$, due to (3.18). Now, recalling (3.13), we have

$$|(\mathbf{i} \cup \mathbf{j} \cup \mathbf{k}) \setminus \mathbf{s}| = |\mathbf{i} \cup \mathbf{j} \cup \mathbf{k}| - |\mathbf{s}| = (2d - w) - (d - 2w + r) = d + w - r.$$

Hence, for fixed $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$, we have $\binom{d+w-r}{2w-r}$ choices for $\mathbf{l}$.

3.5. **Bounding the off-diagonal contribution by a binomial sum.**

We can now return to our bound (3.11) on the off-diagonal contribution. We can rewrite it as follows:

$$S_{\text{off}} \le \sum_{w,v,r} K^{3w/2} \sum_{\mathbf{i} \in \mathbf{I}} \sum_{\mathbf{j} \in \mathbf{J}(\mathbf{i})} \sum_{\mathbf{k} \in \mathbf{K}(\mathbf{i},\mathbf{j})} \sum_{\mathbf{l} \in \mathbf{L}(\mathbf{i},\mathbf{j},\mathbf{k})} |A_{\mathbf{ij}}||A_{\mathbf{kl}}|. \tag{3.19}$$

The first sum is over all realizable v, w, and r, and the rest of the sums are over all possible choices for $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ and $\mathbf{l}$ that conform to the given v, w and r per (3.13). Thus, for instance, $\mathbf{L}(\mathbf{i},\mathbf{j},\mathbf{k})$ consists of all possible choices for $\mathbf{l}$ given $\mathbf{i}$,$\mathbf{j}$ and $\mathbf{k}$. We observed various bounds on realizable v, w and r in Section 3.3, and we computed the cardinalities of the sets $\mathbf{I}$, $\mathbf{J}(\mathbf{i})$, $\mathbf{K}(\mathbf{i},\mathbf{j})$ and $\mathbf{L}(\mathbf{i},\mathbf{j},\mathbf{k})$ in Section 3.4. This knowledge will help us to bound the five-fold sum in (3.19).

In order to do this, rewrite (3.19) as follows:

$$S_{\text{off}} \le \sum_{w,v,r} K^{3w/2} \sum_{\mathbf{i} \in \mathbf{I}} \sum_{\mathbf{j} \in \mathbf{J}(\mathbf{i})} |A_{\mathbf{ij}}| \sum_{\mathbf{k} \in \mathbf{K}(\mathbf{i},\mathbf{j})} \sum_{\mathbf{l} \in \mathbf{L}(\mathbf{i},\mathbf{j},\mathbf{k})} |A_{\mathbf{kl}}|.$$

Note that $|A_{\mathbf{kl}}| \le \|A\| = 1$ for all $\mathbf{k}$ and $\mathbf{l}$, and

$$\sum_{\mathbf{j} \in \mathbf{J}(\mathbf{i})} |A_{\mathbf{ij}}| \le |\mathbf{J}(\mathbf{i})|^{1/2} \Big(\sum_{\mathbf{j} \in \mathbf{J}(\mathbf{i})} A_{\mathbf{ij}}^2 \Big)^{1/2} \le |\mathbf{J}(\mathbf{i})|^{1/2} \|A\| = |\mathbf{J}(\mathbf{i})|^{1/2}.$$

Thus

$$S_{\text{off}} \le \sum_{w,v,r} K^{3w/2} |\mathbf{I}| \cdot \max_{\mathbf{i}} |\mathbf{J}(\mathbf{i})|^{1/2} \cdot \max_{\mathbf{i},\mathbf{j}} |\mathbf{K}(\mathbf{i},\mathbf{j})| \cdot \max_{\mathbf{i},\mathbf{j},\mathbf{k}} |\mathbf{L}(\mathbf{i},\mathbf{j},\mathbf{k})|.$$

Now we can use the bounds we proved in Section 3.4 on the cardinalities of sets $\mathbf{I}$, $\mathbf{J}(\mathbf{i})$, $\mathbf{K}(\mathbf{i},\mathbf{j})$ and $\mathbf{L}(\mathbf{i},\mathbf{j},\mathbf{k})$, which are the number of choices for $\mathbf{i}$, for $\mathbf{j}$ given $\mathbf{i}$, for $\mathbf{k}$ given $\mathbf{i}$, $\mathbf{j}$, and for $\mathbf{l}$ given

[5] Since the number of indices is non-negative, this provides an alternative proof of the bound $r \le 2w$ in Lemma 3.2.

i, **j**, **k**. We obtain

$$S_{\rm off} \le \sum_{w,v,r} K^{3w/2} \binom{n}{d} \binom{d}{v}^{1/2} \binom{n-d}{d-v}^{1/2} \binom{v}{r} \binom{n-(2d-v)}{v-w} \binom{2(d-v)}{d-r-(v-w)} \binom{d+w-r}{2w-r}$$

$$\le \binom{n}{d} \sum_{w,v,r} K^{3w/2} B_1 B_2 B_3 B_4 B_5 B_6, \tag{3.20}$$

where $B_m = B_m(n,d,w,v,r)$ denote the corresponding factors in this expression; for example $B_2 = \binom{n-d}{d-v}^{1/2}$.

3.6. **The terms of the binomial sum.** Let us observe a few bounds on the factors B_m. First,

$$B_5 \le 2^{2(d-v)} \tag{3.21}$$

due to the inequality $\binom{m}{k} \le 2^m$.

Next, since $v \le d+w-r$ by Lemma 3.3, we have $B_3 = \binom{v}{r} \le \binom{d+w-r}{r}$. Combining this with $B_6 = \binom{d+w-r}{2w-r}$, we get

$$B_3 B_6 \le \binom{d+w-r}{r} \binom{d+w-r}{2w-r} \le \binom{d+w-r}{w}^2.$$

Here we used the log-concavity property of binomial coefficients, see Lemma 3.5 in the appendix. Furthermore, we have $w \le d$ by Lemma 3.1 and $r \ge 0$, so

$$B_3 B_6 \le \binom{2d}{w}^2 \le (2ed)^{2w}, \tag{3.22}$$

where we used an elementary bound from Lemma 3.4 in the last step.

Next, using the decay of the binomial coefficients (Lemma 3.6), we get

$$B_4 \le \binom{n-(2d-v)}{v-w} \le \left(\frac{v}{n-2d+1}\right)^w \binom{n-(2d-v)}{v}.$$

Now recall that $v \le d$ (Lemma 3.1) and note that our assumption on d with a sufficiently small constant c implies $d \le n/4$. Thus

$$B_4 \le \left(\frac{2d}{n}\right)^w \binom{n-(2d-v)}{v}.$$

This expression can be conveniently combined with $B_2^2 = \binom{n-d}{d-v}$, since

$$B_2^2 B_4 \le \left(\frac{2d}{n}\right)^w \binom{n-d}{d-v} \binom{n-(2d-v)}{v} = \left(\frac{2d}{n}\right)^w \binom{d}{v} \binom{n-d}{d},$$

The last identity can be easily checked by expressing the binomial coefficients in terms of factorials. This expression in turn can be conveniently combined with $B_1 = \binom{d}{v}^{1/2}$, and we get

$$B_1 B_2 B_4 = B_1 \cdot \frac{B_2^2 B_4}{B_2} = \left(\frac{2d}{n}\right)^w \binom{n-d}{d} \cdot \frac{\binom{d}{v}^{3/2}}{\binom{n-d}{d-v}^{1/2}}. \tag{3.23}$$

Now, using the elementary binomial bounds (Lemma 3.4), we obtain

$$\frac{\binom{d}{v}^{3/2}}{\binom{n-d}{d-v}^{1/2}} = \frac{\binom{d}{d-v}^{3/2}}{\binom{n-d}{d-v}^{1/2}} \le \left(\frac{e^{3/2}d^{3/2}}{(d-v)(n-d)^{1/2}}\right)^{d-v} \le \left(\frac{C_1 d^{3/2}}{n^{1/2}}\right)^{d-v}.$$

In the last step we used that $d-v \ge 1$ by Lemma 3.1 and that $d \le n/2$, which follows from our assumption on d if the constant c is chosen sufficiently small. Recall that by C_1, C_2, etc. we denote suitable absolute constants. Returning to (3.23), we have shown that

$$B_1 B_2 B_4 \le \left(\frac{2d}{n}\right)^w \binom{n}{d} \left(\frac{C_1 d^{3/2}}{n^{1/2}}\right)^{d-v}. \tag{3.24}$$

3.7. **The final bound on the off-diagonal contribution.** We can now combine our bounds (3.21), (3.22) and (3.24) on B_i and put them into (3.20). We obtain

$$S_{\text{off}} \le \binom{n}{d} \sum_{w,v,r} K^{3w/2} B_5 \cdot B_3 B_6 \cdot B_1 B_2 B_4 \le \binom{n}{d}^2 \sum_{w,v,r} \left(\frac{C_2 d^3 K^{3/2}}{n}\right)^w \left(\frac{C_3 d^{3/2}}{n^{1/2}}\right)^{d-v}.$$

Recall from Lemma 3.2 that $0 \le r \le 2w$, thus the sum over r includes at most $2w+1$ terms. Similarly, Lemma 3.1 determines the ranges for the other two sums, namely $0 \le w, v \le d-1$. Hence

$$S_{\text{off}} \le \binom{n}{d}^2 \sum_{w=0}^{d-1} (2w+1) \left(\frac{C_2 d^3 K^{3/2}}{n}\right)^w \cdot \sum_{v=0}^{d-1} \left(\frac{C_3 d^{3/2}}{n^{1/2}}\right)^{d-v}. \tag{3.25}$$

The sums over w and v in the right hand side of (3.25) can be easily estimated. To handle the sum over w, we can use the identity $\sum_{k=0}^{\infty} k z^k = z/(1-z)^2$, which is valid for all $z \in (0,1)$. Thus, the sum over w is bounded by an absolute constant, as long as $C_2 d^3 K^{3/2}/n \le 1/2$. The latter restriction holds by our assumption on d with a sufficiently small constant c.

Similarly, the sum over v in the right hand side of (3.25) is a partial sum of a geometric series. It is dominated by the leading term, i.e. the term where $v = d-1$. Hence this sum is bounded by $C_4 d^{3/2}/n^{1/2}$, as long as $C_3 d^{3/2}/n^{1/2} \le 1/2$. The latter restriction holds by our assumption on d with a sufficiently small constant c.

Summarizing, we obtained the following bound on the off-diagonal contribution (3.5):

$$S_{\text{off}} \lesssim \binom{n}{d}^2 \frac{d^{3/2}}{n^{1/2}}.$$

Combining this with the bound (3.4) on the diagonal contribution and plugging into (3.2), we conclude that

$$\mathbb{E}\left[|\boldsymbol{x}^\mathsf{T} A \boldsymbol{x} - \operatorname{tr} A|^2\right] \lesssim \binom{n}{d}^2 \cdot \frac{Kd^2}{n} + \binom{n}{d}^2 \frac{d^{3/2}}{n^{1/2}} \lesssim \binom{n}{d}^2 \cdot \frac{K^{3/4} d^{3/2}}{n^{1/2}}.$$

In the last step, we used the assumption that $d \lesssim K^{-1/2} n^{1/3}$. The proof of Theorem 1.8 is complete. $\square$

Appendix. Elementary bounds on binomial coefficients

Here we record some bounds on binomial coefficients used throughout the paper.

Lemma 3.4 (see e.g. Exercise 0.0.5 in [45])**.** *For any integers $1 \le d \le n$, we have:*

$$\left(\frac{n}{d}\right)^d \le \binom{n}{d} \le \binom{n}{\le d} \le \left(\frac{en}{d}\right)^d.$$

Lemma 3.5 (Log-concavity of binomial coefficients)**.** *We have*

$$\binom{a}{b-c}\binom{a}{b+c} \le \binom{a}{b}^2.$$

for all positive integers a, b and c for which the binomial coefficients are defined.

Proof. Expressing the binomial coefficients in terms of factorials, we have

$$\frac{\binom{a}{b-c}\binom{a}{b+c}}{\binom{a}{b}^2} = \frac{b!/(b-c)!}{(b+c)!/b!} \cdot \frac{(a-b)!/(a-b-c)!}{(a-b+c)!/(a-b)!}$$

Examining the first fraction in the right hand side, we find that both the numerator and denominator consist of c terms. Each term in the numerator is bounded by the corresponding terms in the denominator. Thus the fraction is bounded by 1. We argue similarly for the second fraction, and thus the entire quantity is bounded by 1. □

Lemma 3.6 (Decay of binomial coefficients)**.** *For any positive integers $s \le t \le m$, we have*

$$\binom{m}{t-s} \le \left(\frac{t}{m-t+1}\right)^s \binom{m}{t}.$$

Proof. The definition of binomial coefficients gives

$$\frac{\binom{m}{t-s}}{\binom{m}{t}} = \frac{t(t-1)\cdots(t-s+1)}{(m-t+s)(m-t+s-1)\cdots(m-t+1)} \le \frac{t^s}{(m-t+1)^s}.$$

□

Lemma 3.7 (Stability of binomial coefficients)**.** *For any positive integers m, p and $t \le m$, we have*

$$\binom{m+p}{t} \le (1+\delta)\binom{m}{t} \quad \textit{where } \delta := \frac{2tp}{m+1-t},$$

as long as $\delta \le 1/2$.

Proof. The definition of binomial coefficients gives

$$\frac{\binom{m+p}{t}}{\binom{m}{t}} = \prod_{k=1}^{p}\left(1 + \frac{t}{m-t+k}\right) \le \left(1 + \frac{t}{m-t+1}\right)^p.$$

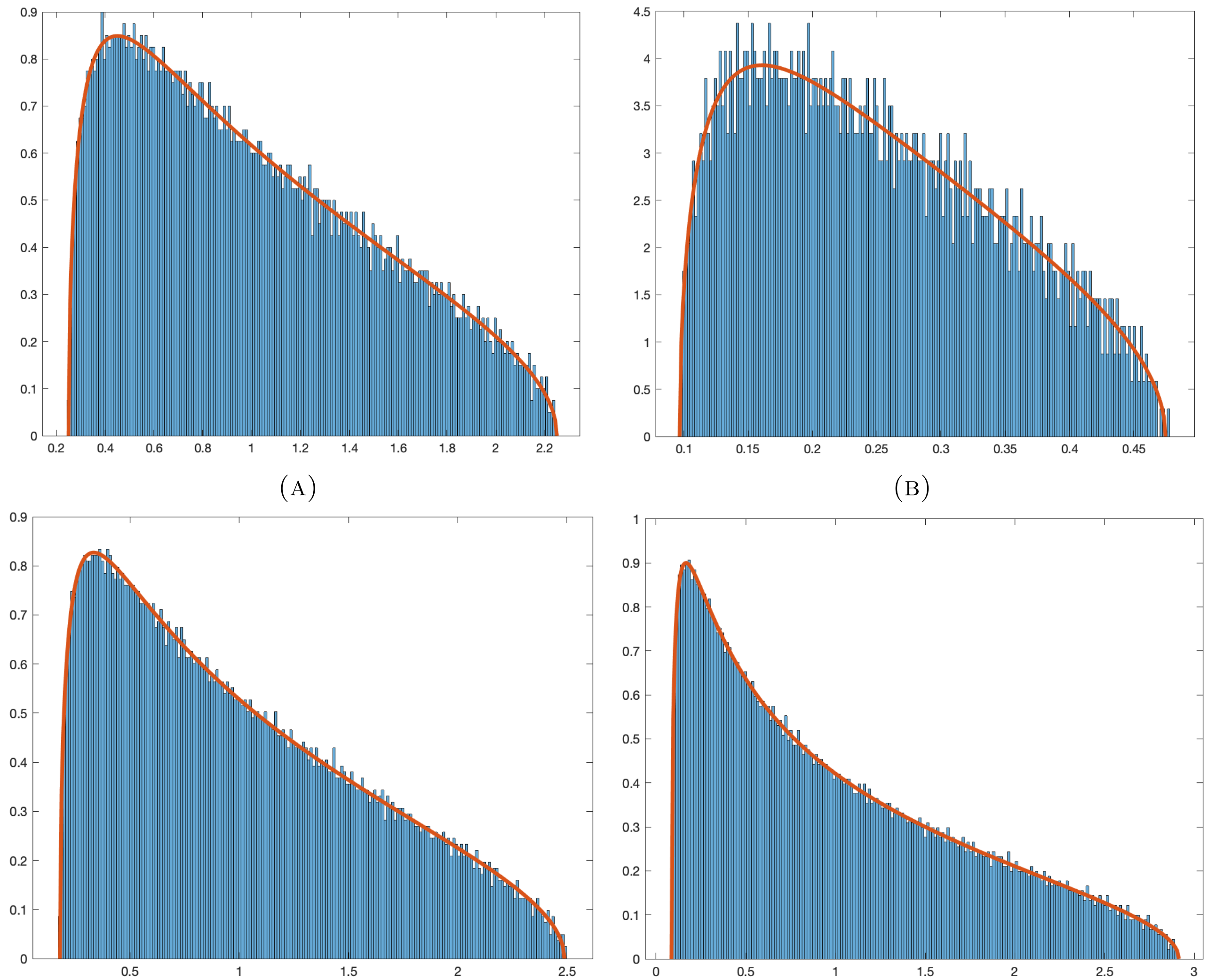
Now use the bound $(1+\epsilon)^p \le e^{\epsilon p} \le 1 + 2\epsilon p$, which holds as long as $\epsilon p \in [0, 1]$. □

4. Numerical Experiments

We present a few numerical experiments to verify that the empirical spectral densities for the block-independent model and the random tensor model tend to the Marchenko-Pastur laws. In all of our tests, the numerical results are computed from a single realization, i.e. we did not average over multiple trials.

Block-independent model experiments: In Figure 1 we show the empirical spectral densities for four experiments of block-independent matrices; in each case, they align very well with the corresponding Marchenko-Pastur density. In Figure 1a, the columns of $X \in \mathbb{R}^{4000\times 16000}$ consist of $n = 2000$ blocks, each of length $d = 2$ where the first entry of the block is $z \sim N(0,1)$ and the second entry is $\frac{1}{\sqrt{2}}(z^2 - 1)$. Thus the second entry is completely determined via a formula of the first entry. While this matrix has half the amount of randomness as an i.i.d. matrix of the same size, it still follows the same limiting distribution as the i.i.d. matrix. We see the densities match up very well even for these relatively small sized matrices. In Figure 1b, the columns of $X \in \mathbb{R}^{1800\times 12600}$ consist of $n = 600$ blocks each of length $d = 3$ where the first and second entry of the block are $\pm\frac{1}{2}$ each with probability $\frac{1}{2}$ and the third entry is a shifted XOR of the first and second (i.e. the third entry is $\frac{1}{2}$ if the first and second entries have opposite signs and it is $-\frac{1}{2}$ if the first and second entries have the same sign). In this case the variance of the entries is $\frac{1}{4}$, so it matches up with Marchenko-Pastur density with covariance matrix $\Sigma = \frac{1}{4}I$ and $\lambda = \frac{1}{7}$. In Figure 1c, the columns of matrix $X \in \mathbb{R}^{7000\times 21000}$ have $n = 10$ blocks, where each block is length $d = 700$ and is of the form $\pm\sqrt{d}e_i$ for i selected uniformly from $[d]$, where $\{e_i\}_{i=1}^d \in \mathbb{R}^d$ are the standard basis vectors in $\mathbb{R}^d$. This example shows that with the exchangeability criteria, it is possible for $n \ll d$. Additionally, we see the two densities agree very well, despite only having $n = 10$ blocks. Similar to Figure 1c, in Figure 1d the columns of matrix $X \in \mathbb{R}^{6400\times 12800}$ have $n = 80$ blocks, where each block is length $d = 80$ and is of the form $\pm\sqrt{d}e_i$ for i selected uniformly from $[d]$. These figures and other experiments together suggest that having $n \geq 10$ and dimensions in the low thousands is enough for the empirical spectral density of a block-independent model matrix to align quite well with the corresponding Marchenko-Pastur density.

Random tensor model experiments: In Figures 2 and 3, we look at vectorized 2-tensors and 3-tensors ($d = 2$ and $d = 3$ respectively). We see that the fourth moment of the entries appears to be important for the speed of convergence as $n \to \infty$. For both the 2-tensors and 3-tensors we consider three types of entries in the vector that we will tensor with itself: 1) the entries are Bernoulli ± 1 each with probability half - these entries have fourth moment of 1; 2) the entries are Uniform on $[-\sqrt{3}, \sqrt{3}]$ - these entries have fourth moment of $\frac{9}{5}$; 3) the entries are standard normal - these entries have fourth moment of 3. In Figure 2 we compare the the empirical spectral density for 2-tensors with the corresponding Marchenko-Pastur density using $n = 145$. We see that the two densities match up quite well, and match up better when the entries had smaller fourth moments. We do the same experiments for 3-tensors in Figure 3 except now using $n = 45$, since $n = 145$ is too computationally costly as it would have $\binom{145}{3} \approx 500{,}000$ rows. We see that the two densities match up quite well for the Bernoulli entry case, not very well for the uniform entry case, and very poorly for the standard normal case. These figures suggest there may even be a different limiting law for small values of n. Testing $n = 100$ does show (Figure 4) that the empirical densities are getting closer to the Marchenko-Pastur density as n increases. These experiments show that while the limiting density does tend to the the Marchenko-Pastur density, they do not align very well for small values of n and the rate of convergence likely depends upon the largest fourth moment of the random vector.



(A) (B) (C) (D)

Figure 1. The Marchenko-Pastur density (red curve) vs. empirical spectral density for block-independent matrices described in Section 4.

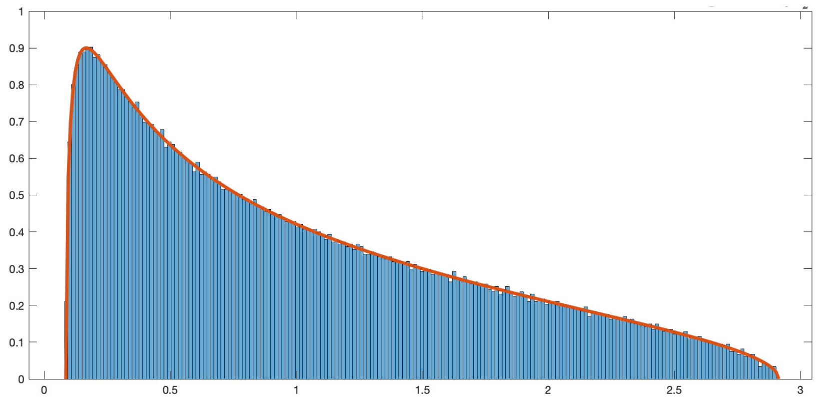
References

[1] R. Adamczak, *Logarithmic Sobolev inequalities and concentration of measure for convex functions and polynomial chaoses,* Bull. Pol. Acad. Sci. Math. 53 (2005), 221–238.

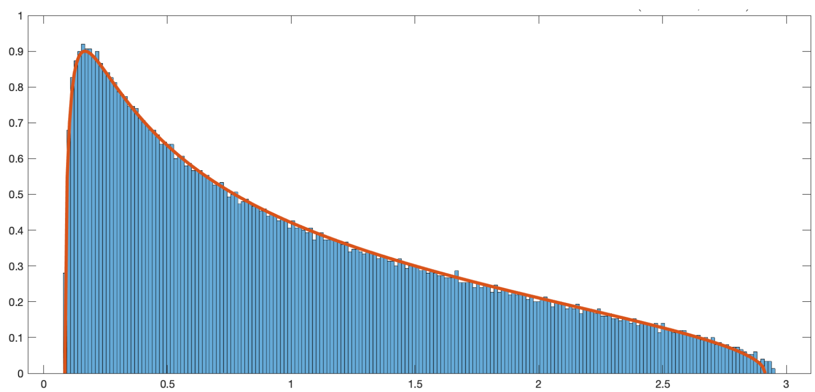
[2] R. Adamczak, *On the Marchenko-Pastur and circular law for some classes of random matrices with dependent entries,* Electron. J. Prob. 16 (2011), no. 37, 1068-1095.

[3] R. Adamczak, R. Latala, *Tail and moment estimates for chaoses generated by symmetric random variables with logarithmically concave tails,* Ann. Inst. Henri Poincaré Probab. Stat. 48 (2012), 1103–1136.

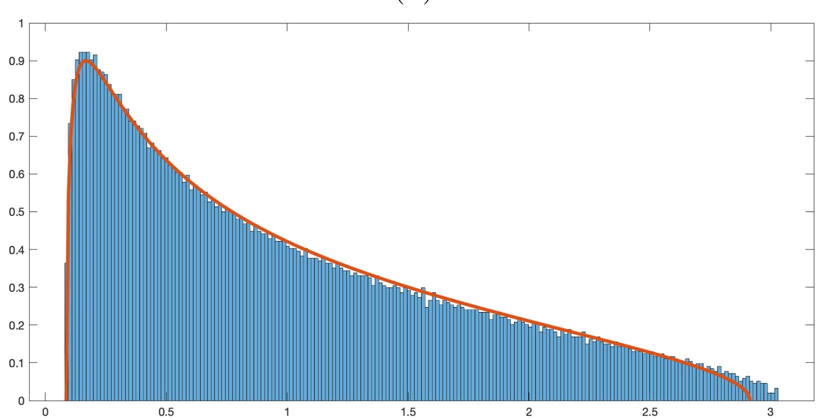
[4] R. Adamczak, P. Wolff, *Concentration inequalities for non-Lipschitz functions with bounded derivatives of higher order,* Probability Theory and Related Fields 162 (2015), 531–586.



(A)



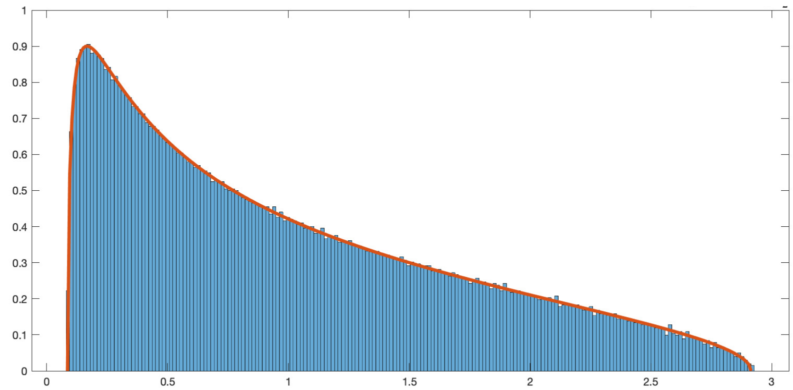
(B)



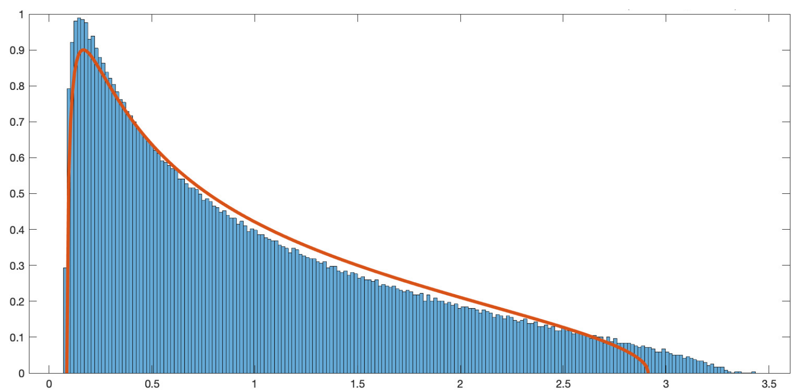
(C)

FIGURE 2. The Marchenko-Pastur density (red curve) vs. empirical spectral density for matrices in $\mathbb{R}^{\binom{145}{2}\times 2\binom{145}{2}}$ whose columns are random 2-tensors as described in Section 4.

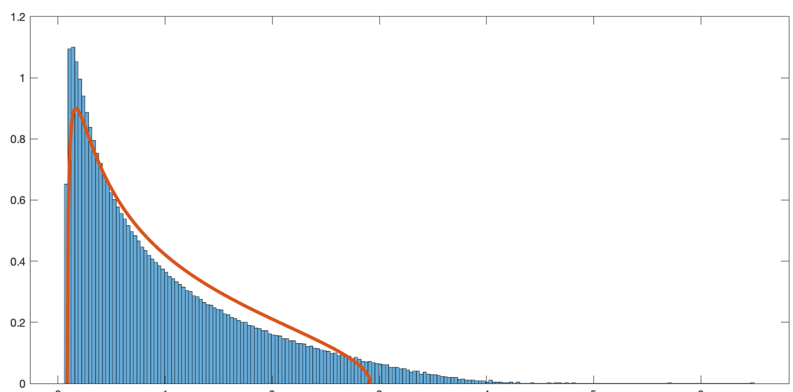
[5] A. Anandkumar, R. Ge, D. Hsu, S. Kakade, M. Telgarsky, *Tensor decompositions for learning latent variable models,* The Journal of Machine Learning Research 15 (2014), 2773–2832.
[6] G. Aubrun, *Random points in the unit ball of l_p^n,* Positivity 10 (2006), no. 4, 755–759.



(A)



(B)



(C)

FIGURE 3. The Marchenko-Pastur density (red curve) vs. empirical spectral density for matrices in $\mathbb{R}^{\binom{45}{3}\times 2\binom{45}{3}}$ whose columns are random 3-tensors as described in Section 4.

[7] A. Auffinger, G. Ben Arous, J. Cerny, *Random matrices and complexity of spin glasses,* Communications on Pure and Applied Mathematics 66 (2013), 165–201.

[8] Z. D. Bai, *Methodologies in spectral analysis of large-dimensional random matrices, a review*, Statistica Sinica 9 (1999), 611–677.

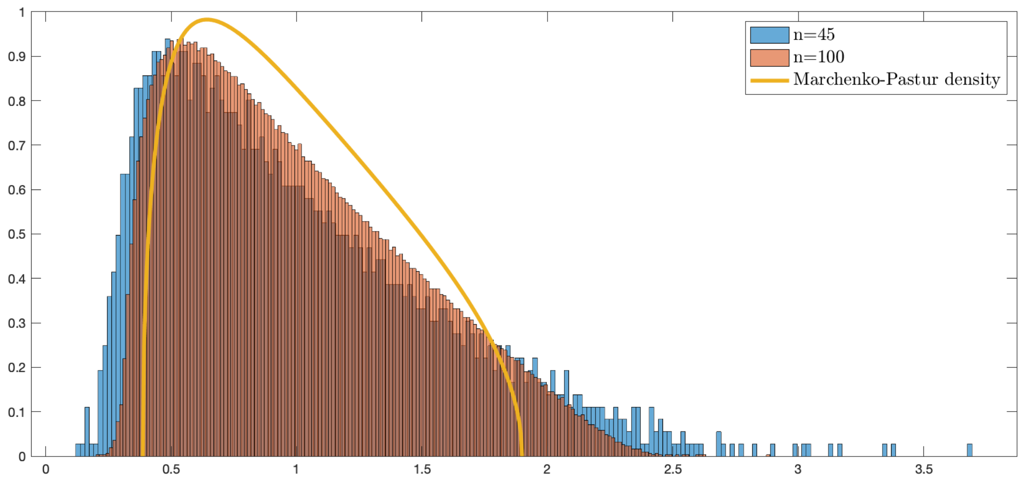


FIGURE 4. Empirical spectral density of $\frac{1}{\binom{n}{3}}XX^T$, where columns of $X^T \in \mathbb{R}^{\binom{n}{3} \times \frac{1}{7}\binom{n}{3}}$ are 3-tensors of a random vector in $\mathbb{R}^n$ with entries uniform on $[-\sqrt{3}, \sqrt{3}]$ as described in Section 4.

[9] Z. D. Bai and J. W. Silverstein, *On the empirical distribution of the eigenvalues of a class of large dimensional random matrices,* J. of Multivariate Analysis 54 (1995), no. 2, 175–192.

[10] Z. D. Bai and J. W. Silverstein, *Spectral analysis of large dimensional random matrices.* 2nd ed., Springer, 2010.

[11] Z. Bai and W. Zhou, *Large sample covariance matrices without independence structures in columns,* Statistica Sinica 18 (2008), 425–442.

[12] P. Baldi, R. Vershynin, *Polynomial threshold functions, hyperplane arrangements, and random tensors,* SIAM Journal on Mathematics of Data Science, to appear.

[13] D. Banerjee and A. Bose, *Bulk behavior of some patterned block matrices,* Indian J. Pure Appl. Math. 47 (2016), no. 2, 273–289.

[14] G. Ben Arous, S. Mei, A. Montanari, M. Nica, *The landscape of the spiked tensor model,* Communications on Pure and Applied Mathematics 72 (2019), 2282–2330.

[15] W. Bryc, A. Dembo and T. Jiang, *Spectral measure of large random Hankel, Markov and Toeplitz matrices,* Ann. Prob. 34 (2006), no. 1, 1–38.

[16] W.-K. Chen, *Phase transition in the spiked random tensor with Rademacher prior,* The Annals of Statistics 47 (2019), 2734–2756.

[17] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications.* Cambridge University Press, 2011.

[18] E. Dobriban, *Efficient computation of limit spectra of sample covariance matrices,* Random Matrices: Theory and Applications 04 (2015), no. 4, 1550019–1550055.

[19] R. Ge, F. Huang, C. Jin, Y. Yuan, *Escaping from saddle points – online stochastic gradient for tensor decomposition,* COLT 2015 (Conference on Learning Theory), 797–842.

[20] F. Götze, H. Sambale, A. Sinulis, *Concentration inequalities for polynomials in α-sub-exponential random variables,* preprint (2019).

[21] F. Götze, A. Tikhomirov. *Limit theorems for spectra of positive random matrices under dependence,* Zap. Nauchn. Sem. S.-Petersburg. Otdel. Mat. Inst. Steklov. (POMI), Vol. 311 (2004), Veroyatn. i Stat. 7, 92–123, 299.

[22] F. Götze, A. Tikhomirov. *Limit theorems for spectra of random matrices with martingale structure,* Stein's Method and Applications, Singapore Univ. Press (2005), 181–195.

[23] N. El Karoui, *Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond,* Ann. Appl. Probab. 19 (2009), no.6, 2362–2405.

[24] D. Ghoshdastidar, A. Dukkipati, *Consistency of spectral hypergraph partitioning under planted partition model,* The Annals of Statistics 45 (2017), 289–315.
[25] U. Greander and J. W. Silverstein, *Spectral analysis of networks with random topologies,* SIAM J. Appl. Math. 32 (1977), 499–519.
[26] P. Jain, S. Oh, *Provable tensor factorization with missing data,* NIPS 2014 (Advances in Neural Information Processing Systems), 1431–1439.
[27] D. Jonnson, *Some limit theorems for the eigenvalues of a sample covariance matrix,* J. Multivariate Anal. 12 (1982), 1–38.
[28] C. Kim, A. S. Bandeira, M. X. Goemans, *Community detection in hypergraphs, spiked tensor models, and sum-of-squares,* SampTA 2017 (International Conference on Sampling Theory and Applications), 124–128.
[29] L. Laloux, P. Cizeau, M. Potters and J-P. Bouchaud, *Random matrix theory and financial correlations,* International Journal of Theoretical and Applied Finance 1 (2000), no. 03, pp.391–397.
[30] R. Latala, *Estimates of moments and tails of Gaussian chaoses,* The Annals of Probability 34 (2006), 2315–2331.
[31] R. Latala, R. Lochowski, *Moment and tail estimates for multidimensional chaos generated by positive random variables with logarithmically concave tails,* Stochastic inequalities and applications (2003), 77–92.
[32] J. Lehec, *Moments of the Gaussian chaos,* Seminaire de Probabilites XLIII, 327–340, Lecture Notes in Math. (2006), Springer, Berlin, 2011.
[33] J. Lei, K. Chen, B. Lynch, *Consistent community detection in multi-layer network data,* Biometrika, to appear (2019).
[34] P. Loubaton, *On the almost sure location of the singular values of certain Gaussian block-Hankel large random matrices,* J. Theoretical Probab., submitted.
[35] V. A. Marchenko and L. A. Pastur, *Distribution of eigenvalues for some sets of random matrices,* Math USSR Sbornik 1 (1967), 457–483.
[36] A. Montanari, N. Sun, *Spectral algorithms for tensor completion,* Communications on Pure and Applied Mathematics 71 (2018), 2381–2425.
[37] N. H. Nguyen, P. Drineas, T. D. Tran, *Tensor sparsification via a bound on the spectral norm of random tensors,* Information and Inference: A Journal of the IMA 4(2015), 195–229.
[38] S. O'Rourke, *A note on the Marchenko-Pastur law for a class of random matrices with dependent entries,* Electron. Commun. Probab. 17 (2012), no. 28 ,1–13.
[39] A. Pajor and L. Pastur, *On the limiting empirical measure of eigenvalues of the sum of rank one matrices with log-concave distribution,* Studia Math. 195 (2009), no. 1, 11–29.
[40] V. Plerou, P. Gopikrishnan, B. Rosenow, L. Amaral, T. Guhr and H. Stanley, *Random matrix approach to cross correlations in financial data,* Phys. Rev. E 64 (2002), 66126-66144.
[41] E. Richard, A. Montanari, *A statistical model for tensor PCA,* NIPS 2014 (Advances in Neural Information Processing Systems), 2897–2905.
[42] M. Rudelson, R. Vershynin, *Hanson-Wright inequality and sub-gaussian concentration,* Electronic Communications in Probability 18 (2013), 1–9.
[43] J. Silverstein, *Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices,* J. Multivariate Analysis 55 (1995), no. 2, 331–339.
[44] T. Tao, *Topics in random matrix theory.* Graduate Studies in Mathematics, 132. American Mathematical Society, Providence, RI, 2012.
[45] R. Vershynin, *High-dimensional probability: an introduction with applications in data science.* Cambridge University Press, 2018.
[46] R. Vershynin, Concentration inequalities for random tensors, submitted (2019).
[47] K. W. Wachter, *The strong limits of random matrix spectra for sample matrices of independent elements,* Ann. Probab. 6 (1978), 1-18.
[48] Y. Wang, H.-Y. Tung, A. J. Smola, A. Anandkumar, *Fast and guaranteed tensor decomposition via sketching,* NIPS 2015 (Advances in Neural Information Processing Systems), 991–999.

[49] P. Yaskov, *A short proof of the Marchenko-Pastur theorem,* C. R. Math. Acad. Sci. Paris Ser. I, 354 (2016), 319–322.

[50] P. Yaskov, *Necessary and sufficient conditions for the Marchenko-Pastur theorem,* Electron. Commun. Probab. 21 (2016), 73-80.

[51] J. Yao, *A note on a Marchenko-Pastur type theorem for time series,* Statist. and Probab. Letters 82 (2012), 20-28.

[52] Y. Q. Yin and P. R. Krishnaiah, *Limit theorems for the eigenvalues of product of large-dimensional random matrices when the underlying distribution is isotropic,* Teor. Veroyatnost. i Primenen. 31 (1986), 394-398.

[53] Y. Q. Yin, *Limiting spectral distribution for a class of random matrices,* J. Multivariate Anal. 20 (1986), 50-68.

[54] Z. Zhou, Y. Zhu, *Sparse random tensors: concentration, regularization and applications,* submitted (2019).

Department of Mathematics, University of California, Irvine, Irvine, CA. 92697
E-mail address: jabryson@uci.edu
E-mail address: rvershyn@uci.edu
E-mail address: zhao@uci.edu